30 More than seventy years of probabilistic phonology

Janet B. Pierrehumbert

Abstract: The idea that phonology is probabilistic goes back to Pāṇini, and the classic distinction between accidental and systematic gaps in the lexicon is implicitly probabilisitic. This idea took on new vigour just over 70 years ago with Shannon's work on information theory. It is now supported by a wide variety of studies concerning all levels of representation in the theory of sound structure. Models that use probabilities have had striking successes in capturing how people learn the sound patterns of their language from experience, and deploy their knowledge productively and adaptively. However, people's patterns in production often deviate systematically from the statistical patterns in the input they receive. Cognitive biases, propensities for regularity and structure, and social influences on attention and memory, all play important roles.

30.1 Introduction

Human languages have extremely large lexicons, compared to animal communication systems. These lexicons are made possible by the phonological principle, according to which a relatively small number inventory of elements of sound structure that are not meaningful in themselves can be recombined in many different ways to create a much larger number of complex forms that are associated with meanings. This

means that any given word or phrase can be analysed either at the phonological level or at the semantic level, according to the concept of 'duality of patterning' developed in Hockett (1958, 1960); see Ladd, this volume, for further discussion. The inventory of phonological elements is language-specific, as are the constraints on their combinations. However, there are striking regularities in the possible relationships amongst elements and the systems of constraints that are found across languages. These regularities point to commonalities not only in the phonetic grounding of phonology in human articulatory and perceptual capabilities, but also in the cognitive system that allows individuals to acquire and use a phonological system from experience with language. In this chapter, I will review the history of an important claim about this cognitive system, namely the claim that statistics play a central role: the phonological grammar is acquired by a process of statistical inference over linguistic events of different frequencies, and furthermore the resulting mental representations incorporate probabilities in some manner. Note that following standard usage, I will use the term 'frequency' for how often something happens, and the word 'probability' for the likelihood that it would happen, in comparison to alternative outcomes. Frequency is reported as a count, whereas probability is reported as a number between zero and one, obtained by dividing the frequency by the total count for all the different alternatives.

The claim that phonology is probabilistic has a long history, but a turning point in this approach was the development of information theory just over 70 years ago in Shannon (1948). Information theory provides a

rigorous foundation for formalizing and evaluating this claim; hence the title of this chapter. The claim follows—indirectly—from the fact that any empirically observed lexicon is a sparse sample of the forms that adult speakers would accept and would be able to produce, encode, and remember. Phonologists as language scientists aim for synoptic phonological grammars of individual languages. That is, by providing a compressed, abstract, and general description of the word forms in a language, the grammars should encompass not only the words that were observed, but also words that might exist, but weren't observed. Like other scientists, we are inspired to evaluate the success of our theories by testing them against previously unseen data. But furthermore, ordinary speakers resemble scientists because their implicit knowledge is itself abstract, general, and predictive, as revealed in the ability to encode, remember, and create previously unseen word forms. In addition to borrowings from other languages and novel proper names, these include novel combinations of morphemes, word blends, technical terms, and slang words. Thus, the acquisition of phonology resembles the process of constructing a scientific theory by a process of inference from patterns in the data. As with many other examples of scientific inference, a probabilistic model has access to information that is discarded in models without frequency information, and it can use this information to succeed in making predictions. More precisely, it can exploit empirical observations more optimally than a model using only *Boolean logic*, in which the value of any variable or formula can only be 1 (*true*) or 0 (*false*). Because a Boolean model assigns the value of 1 (true) to anything that was observed—no matter how often or rarely it was

observed—it throws away information. It is, for example, unable to predict that novel combinations of things that are individually frequent are much more likely to appear in the future than novel combinations of things that are individually rare. Insofar as people behave optimally and use all available information to learn phonological generalizations, we hypothesize that they use probabilistic information.

Implicit knowledge of phonological generalizations is revealed in a wide variety of productive or adaptive behaviour. Classically, generative phonology has main concerns. First, is the phonological grammar of any given language capable of enumerating all and only the possible words? This criterion is intrinsically probabilistic because the observed words are considered to be a random sample from a very much larger set of possible words. Thus, the set of words that are well-formed according to the grammar needs encompass all the observed words[1]. Delving in more deeply, the accidental gaps are just the possible words that – entirely by chance – failed to appear in the sample; according to the Oxford English Dictionary, 'accidental' means 'relating to or occurring by chance'. And the concept of chance is exactly what receives a precise construction in the theory of probability. Because accidental gaps occurred by chance, we expect that they might be real words in an alternative universe in which the lexicon was a different random sample of the possible words.

A particularly accessible alternative universe is the one in which we simply ask native speakers to add a few words to their existing lexicon – for

[1] Exceptions to this requirement are sometimes made for a small number of anomalous forms, such as foreign proper names and onomatopoeic expressions.

example, by asking them to adopt nonce forms like /strɪmpi/ and /zgɛmθu/ as slang words in a futuristic novel, or names for new products. If /strɪmpi/ represents an accidental gap in the lexicon, we expect that speakers will accept it readily, and indeed they do (Hay, Pierrehumbert, & Beckman 2004*a*). In contrast to accidental gaps, systematic gaps did not occur by chance and are not expected to be filled. If English speakers reject /zgɛmθu/, this would be evidence that it represents a systematic gap. Thus, the distinction between accidental and systematic gaps is directly connected to the second classic concern of generative phonology, namely well-formedness judgements of nonce word forms. Well-formedness judgements represent the conscious, meta-level, application of generalizations that are otherwise merely implicit. Such judgements have turned out to be gradient in a manner that reflects empirical frequencies.

These two outcomes do not exhaust the sources of information about the cognitive representation of phonology, however. Unconscious interpretative processes are at least as important and illuminating. These include the inferences that allow infants to acquire the phonology of their native language. The processes of encoding speech signals, forming lexical representations for novel words, and perceptually adapting to different interlocutors in adulthood also reveal aspects of the cognitive representation of phonology.

Research on cognitive models of phonology has actively explored the role of probabilistic information for more than seven decades, leading to many different observations and theories. An important dimension along which these theories vary is their claims about the relationship between

empirical frequencies and the cognitive reflexes of frequency. In the extreme, some researchers, such as Labov (1989), claim that humans typically probability-match the input. That is, any variation in the input will be cognitively encoded and matched in subsequent productions. Other researchers have focussed on more general outcomes that indicate the importance of empirical frequencies, even if the outcomes differ significantly from probability matching. Arguing against purely Boolean models of phonology, these researchers have established the existence of gradient and cumulative effects that are correlated with empirical frequencies. However, in some cases the outcomes differ significantly from probability-matching, because they exaggerate, attenuate, or filter empirical frequencies. To explain such outcomes, these researchers propose cognitive mechanisms that have more complex or indirect sensitivity to frequency than a simple probability-matching learning algorithm would have; these proposals will be further discussed in Section 30.6.

In this chapter, we first review some of the concepts and results that provide a foundation for probabilistic theories of phonology. Any theory of phonology needs an ontology, understood as an inventory of entities and relations that play a role in the theory. In a probabilistic theory, these become the entities and relations that may support probabilistic scores. For example, if the phoneme is a unit in the theory, the frequency of the phoneme is available in the probabilistic theory. In Section 30.2, we begin by reviewing the levels of representation and the entities defined at different time scales that play a role in any insightful and predictive theory. In Section 30.3, we develop the relationship between statistical learning and

productivity. Section 30.4 reviews probabilistic models of the mappings between phonology and phonetics (on the one hand) and morphophonology (on the other hand). Section 30.5 provides an overview of experimental results establishing correlations between corpus statistics and various kinds of linguistic behaviour. Finally in Section 30.6, we discuss the extent to which the correlations described in Section 30.5 indicate that language learning is probability-matching, meaning that the likelihoods of different variants are the same in the learner's output as they were the input the learner experienced. We will review some of the most important deviations from probability-matching that have been established, but still conclude that probability matters.

30.2 Probabilities of what?

Since the 19th century work of Baudoin de Courtenay (Koerner 1972; Radwańska-Williams, this volume), it has been acknowledged that language sound structure involves (at least) three levels of representation: phonetics, word-level phonology, and morphophonology. Lakoff (1993) provides a recent and clear exposition of how constraints within levels and between pairs of levels can effectively capture patterns in Mohawk, Lardil, and other languages previously thought to require deep rule ordering. The model would allow five different loci for probabilistic information to be included: the three levels of representation, plus the relations between adjacent levels (relations of morphophonology to word-level phonology, and relations of word-level phonology to phonetics).

The research literature has used all five of these loci. Here are some examples. Probabilistic constraints applying within the morphophonological level are exemplified by the constraints in Arabic and many other languages that disfavour combinations of homorganic consonants in close proximity (McCarthy 1988, 1994; Frisch, Pierrehumbert, & Broe 2004). These are best defined at the morphophonological level because their effects are obscured at the phonological level by cases in which multiple copies of the same consonant fill out word-level structural templates. Within Finnish word-level phonology, there is a dispreference for heavy syllables containing high vowels, and also for light syllables containing low vowels. Extending the framework of Optimality Theory, Anttila & Cho (1998) develop a probabilistic model of how the competition between these two constraints plays out for a subset of the forms for the genitive plural. Some stems may have either a trisyllabic genitive plural form (which meets one constraint) or a disyllabic form (which meets the other constraint). The synchronic variation between these two outcomes is effectively captured through variable constraint ranking. At the phonetic level, there is evidence that infants implicitly learn many statistical properties of the the acoustic-phonetic patterns of their language even before they acquire the lexical inventories that define the phonological level (Werker & Tees 1984). Variable rules, the mainstay of sociophonetics (Sankoff & Labov 1979), typically describe the probabilistic relation between word-level phonology and a phonetic outcome. However, analogous issues arise in the variable relations between the morphophonological and phonological levels (Guy 1991). While the standard formalization of variable rules using logistic

regression was only achieved in the 20th century, the core concepts date back to the ancient Sanskrit grammarian Pāṇini (Cardona 1965); for further discussion see Kiparsky, this volume.

Within any given level, there are also important questions about the interactions of probabilities at different time scales. Some articulators, such as the tongue tip, are much faster than others (such as the jaw or the velum), and some contrasts in sound structure are realized as much faster events than others. For example, the flap in an English word like *putting* is typically around 25 msec, whereas the rise in fundamental frequency for a yes/no question (heard as a rise in pitch) can easily take 40 times longer, spanning many syllables. How should we understand the relationships amongst regularities that can be observed at all these different time scales? Different versions of phonemic theory were ill-suited to integrating information at different times scales because they posited a one single privileged timing unit (the phoneme), mapping to the allophone at the phonetic level. In these theories, the phonological representation resembled 'beads on a string', in which 'beads' (the phonemes or speech segments) were assembled in sequences, and no suprasegmental or subsegmental structures were defined. While it was never claimed that all phonemes had the same duration, phonemes still had a privileged role in explaining how speech is produced in time as a sequence of articulatory actions and perceived as a sequence of contrastive acoustic events. The traditional taxonomic phonemic level was eliminated in Chomsky & Halle (1968), and the systematic phonemic level they proposed is more similar to the

morphophonological level in the traditional theory;[2] the representation manipulated throughout the derivation is a matrix in which the rows are distinctive features and the columns are phoneme-sized bundles of distinctive features. However, this theory retains the assumption that the privileged timing unit is phoneme-sized. While Chomsky & Halle (1968) and others acknowledged that some systematic patterns (such as is the complicated alternating patterns of English stress) appear to be defined at larger time scales, the theoretical account of these patterns coerces them into a Procrustean bed of phoneme-sized timing units.

During the 1970s and 1980s, autosegmental-metrical phonology systematically explored phonological regularities at different time scales, and proposed an ontology in which units larger than the phoneme also play critical roles.[3] Phonological constraints at time scales larger than the phoneme are described with reference to these larger units. Defining these units in the right way allows simple formulations of many constraints that appear complex and unwieldy when described using phoneme sequences or distinctive feature matrices. Widely accepted units include the syllable, the metrical foot, and the intonation phrase. Cross-cutting the hierarchical units of metrical phonology are the domains of important autosegmental constraints, which refer to spans of timing units. Examples include tone spreading (whereby a tone is realized on a sequence of vowels, and not just one vowel), vowel harmony, and constraints on voicing or nasal assimilation at syllable junctures (Goldsmith 1990; Pierrehumbert & Beckman 1988; Ito

[2] For discussion, see Ladd and Dresher & Hall, both this volume.

[3] On autosegmental and metrical phonology, see Kisseberth, this volume.

1988).

Both autosegmental constraints and metrical trees can be treated probabilistically. Autosegmental constraints fundamentally deal with sequential probabilities, and formal approaches all extend Shannon's use of Markov models in analyzing sequences of letters (see below). The articles on Arabic mentioned above all analyze probabilities on autosegmental tiers that have projected away from the vowels that come between the consonants. The tree structures used in the prosodic hierarchy might seem to require a more powerful formalism, such as a stochastic context-free grammar (Charniak 1997). However, extended Markov models as described below have been more used in practice. Examples of this approach include the treatment of lexical stress in Coleman & Pierrehumbert (1997) and the treatment of phrasal intonation in Ostendorf & Ross (1997).

30.3 Statistical learning and productivity

At central goal of generative phonology is to understand how language learners can acquire the general abstract system that enables them to produce and understand novel forms. This means understanding the relationship between learning and productivity. A good point of departure for understanding this relationship is Shannon (1948) (see also Shannon & Weaver 1949), which laid out the foundations of information theory, and was a major influence on the work of Charles F. Hockett, Roman Jakobson, and others (as discussed in Dresher & Hall, this chapter). The paper includes a section on the applicability of Markov processes for describing

English text. We can take English letters as analogues of phonemes, in that they provide a small number of coding symbols that are combined in various orders to specify word forms. Markov processes are examples of finite state grammars, which represent the lowest and most tractable level of what subsequently became known as the Chomsky hierarchy (Chomsky 1956).[4] Finite state grammars capture local constraints on sequences, where 'local' means that the constraints can be defined using a fixed-sized window on the sequences that occur in the language. A sequence of n consecutive elements is referred to as an n-gram. A unigram (or 1-gram) grammar allows only a single element to be visible at once. A bigram (or 2-gram) grammar allows sequences of two elements to be visible at once, and so forth. Taking the last position in the n-gram as the 'current' position, the order n of the grammar thus determines the size of the history that can be taken into account in constraining the current position. In a non-stochastic finite state grammar, a transition is either possible or not. However, in a Markov process, probabilities are assigned to the transitions, capturing the fact that some continuations of any given sequence are likely while others are rare, but still possible.

Having only local constraints, Markov processes provide very attractive correspondences amongst learning, parsing, and generation. The nature of these correspondences is known mathematically, and they are efficiently computable. A grammar can be acquired from a linguistic sample by tabulating all the n-grams of the relevant order in the sample, with their frequencies. This grammar can be used as a parser; given an example of a

[4] On finite state grammers, see Chandlee & Jardine, this volume.

new sequence, it can determine whether or not a new sequence is within the language of the training set at all; and if it is, whether it is more or less likely to occur. The grammar can also be used as a generator, in which case it will produce not only existing sequences but also novel sequences, thus providing predictions about phonological productivity.

Shannon's tutorial examples illustrate the sequences that can be generated for letter and word grammars of various orders as trained on an English corpus. Examples like (1) result if letters are selected at random and are equally probable.

(1) XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYVKCQSGHYD ...

A unigram model that selects letters with the empirically observed frequencies, but without any reference to the context, yields the following example:

(2) OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI ALHENHTTPA ...

(2) improves over (1) chiefly because it uses the letters for vowels much more frequently than (1) does. However, substrings such as NBN in (2) reflect the fact that the contextual constraints are being completely ignored. Using bigram statistics, the results resemble English to some extent.

(3) ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY ACHIN D ILONASIVE ...

With trigram statistics, the results begin to look quite realistic:

(4) IN NO IST LAT WHEY CRATICT FROURE BIRS GROCID PONDENOME OF DEMONSTURES OF THE REPTAGIN IS ...

The comparison of these four models suggests that probabilistic sequential constraints are very important in phonology. The bigram model (3) improves on (1) and (2) because it can capture the alternation of consonants and vowels in core syllables, as well as constraints on syllable contacts. However, it still generates some impossible sequences, such as word-initial CT. Since C can occur before T in English (as in *act*), the bigram grammar will permit it in all positions. The trigram model is capable of allowing CT medially or finally but not word-initially, because it can refer to the word boundaries as if they were phonemes, differentiating #CT sequences (which do not exist in English), from CT# sequences, which do exist. In summary, each n-gram model generates a subset of the sequences that are generated by the next simpler model. The higher the value of n, the more stringent the n-gram model and the more its outputs conform to the actual patterns of English. In technical terms, the *precision* of the model (the fraction of generated forms that are valid) increases as the order increases. The flip side is that the *recall* (the fraction of valid forms that are generated) deteriorates if the order becomes too high. A 4-gram or 5-gram model is incapable of generating sequences such as HANCH and LONT that are clearly accidental gaps in the lexicon. In the limit, the model includes only attested words and has no ability to generate or parse novel words at all. Thus, the best model must strike a balance between being overly simple (and therefore over-generating) versus being overly detailed (and therefore failing to support generalizations). This

observation about n-gram models extends to all models that support the triad of learning, generation, and parsing.

Shannon also shows what happens when ngram models are built over words instead of over letters. (5) is an output from a bigram word level model.

(5) THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER OF THIS POINT IS THEREFORE ...

Because a single word can contain many letters, (5) illustrates how a hierarchy of Markov models can be used to capture regularities at multiple time scales. The approach is not successful as an implementation of syntactic patterns such as *wh*-movement, because the model cannot generalize across cases in which the material intervening between the *wh*-word and its trace varies in length and complexity, a central point of Chomsky (1957). From the point of view of sound structure, however, the n-gram model looks better. There is no problem in reading out (5) as a phonologically well-formed sequence of intonation phrases as in (6).

(6) THE HEAD AND IN FRONTAL ATTACK % ON AN ENGLISH WRITER % THAT THE CHARACTER OF THIS POINT % ...

Although metrical phonology does posit hierarchical structures, it has been argued that this involves a limited number of levels, with the CV skeleton and/or moraic structure at the bottom, and the intonation phrase at the top. Accordingly, it is claimed to lack recursion (Nespor & Vogel 1986; Beckman & Pierrehumbert 1986). Insofar as this claim is correct, it

entails that metrical trees can be handled using multi-level Markov models. Ostendorf & Ross (1997) use this approach in an algorithm for recognizing intonation patterns in recordings of speech. To include word-level metrical structure in building a probabilistic grammar to predict well-formedness judgements, Coleman & Pierrehumbert (1997) simply compile the metrical position into the set of states for a single-level grammar. However, stochastic context-free grammars also support learning, generation, and parsing (Chappelier, Rajman *et al.* 1998). So recursive prosodic grammars would not be fatal to the enterprise of linking production to learning and perception, though in practice, they have been little used in statistical models of phonological acquisition.

30.4 Relating phonology to phonetics and to morphophonology

Phonology can be rather well approximated as a discrete system. Phonetics characterizes physical speech events (whether articulatory gestures or acoustic patterns). Thus, phonological elements categorize a physical space with a large number of dimensions, in much the same way as colour words categorize a continuous multi-dimensional world of the visual spectrum. It is challenging to capture the relationship between such fundamentally disparate levels of representation (a discrete set of coding symbols on the one hand, and a quantitative representation of physical reality on the other hand), and the high dimensionality of the phonetic space creates its own set of challenges. However, a positive feature of the situation is that the

phonetic parameters are physically observable, and the scientific theory of the relationship between articulation and acoustics is very advanced.

After recordings of conversational speech became widely available in the 1960s and 1970s, fine phonetic transcriptions of these recordings launched two important lines of research in probabilistic phonology. The VARBRUL framework developed in Sankoff & Labov (1979) formalized the relationship of phonemes to phonetic realizations as rewrite rules with associated probabilities. For example, the probability distribution for American English /t/ (in dimensions that include the closure, the vocal fold configuration, and the formant transitions) is coarse-grained as the set of alternatives {[t^h], [t], [$t^?$], [ʔ], [ɾ]}. The probabilities for each of these variants, as a function of social characteristics such as gender, age, and class, can be inferred from a labelled corpus. According to usage-based phonology (Bybee 2001; Kapatsinski 2018), phonology is a self-organizing system in which ongoing experience of language continually updates the mental representations, which are characterized as richly detailed memories organized into an associative network. Frequency effects are an intrinsic consequence of this learning mechanism. The original works in usage-based phonology, notably Hooper (1976), shared with the VARBRUL framework the use of phonetic transcription. A major empirical finding was strong correlations between word frequencies and the probabilities of lenited allophonic variants.

Variability in American English /t/ lends itself very well to the VARBRUL approach, because the different variants correspond to distinct clusters in the acoustic-phonetic space. However, for other cases studied in

sociophonetics and usage-based phonology, the variation is systematic, distinct clusters are not apparent to us. Vowel formants, fundamental frequency contours, and rhythmic patterns can all vary in ways that display regular relationships to factors such as prosodic position, gender, or dialect, but without the variation being easily transcribed in fine phonetic notation. For example, Hay, Pierrehumbert, Walker, & LaShell (2015) identify effects of word frequency on vowel formants during a regular sound change in progress in New Zealand that are extremely subtle – just a few Hertz in size – yet still systematic and statistically significant. Similarly, Sanchez, Hay, & Nilson (2015) find that New Zealand vowels have a slightly more Australian realization in the context of discussions about Australia. Attempting to transcribe such patterns with a small inventory of symbols would omit details that are interesting and produced systematically by native speakers. Of course the inventory could in principle be elaborated to any desired level of precision, perhaps using 10 or 100 times as many vowel symbols as the IPA committee would countenance. But if the transcription system is extremely detailed, it becomes impossible to achieve good reliability across transcribers. More importantly, the inventory of any transcription system effectively incorporates assumptions about alternative categorical choices in the cognitive system. For example, while I normally produce *pretty* with a [ɾ], I may categorically decide to use [t^h] to make myself understood to a British speaker. In contrast, for some aspects of speech production, it is more reasonable to assume that the speaker is controlling a continuous variable such as their precision or level of effort, much as they can control their precision or level of effort in other physical tasks.

An alternative theoretical approach does not stop at generating a fine phonetic transcription, but instead associates phonological categories with probability distributions over observable physical parameters, such as the formants that capture vowel quality distinctions. This approach is possible because of the tremendous advances in speech science in the 1940s through the 1970s, including the application of linear systems theory to vowel acoustics (Chiba & Kajiyama 1942; Potter, Kopp, & Green 1947; Fant 1970), van den Berg's theory of vocal fold vibration (Van den Berg 1958), and von Bekesy's work on the mechanisms of the ear, which led to the Nobel Prize in 1961 (Olson, Duifhuis, & Steele 2012). An early example of phonological explanation using these tools is adaptive dispersion theory, which explains some aspects of the typology of vowel systems through an iterated stochastic model of production and perception (Liljencrants & Lindblom 1972). This is an example of a usage-based model, because the mental representation of the vowel system emerges from ongoing experience with speech.

Key assumptions of adaptive dispersion theory are carried over to recent work in exemplar theory (Goldinger 1998; Pierrehumbert 2001; Johnson 2005; Wedel 2012). In perception, the incoming signal is classified on the basis of its location in the phonetic space by deciding which category it is most likely to belong to, a statistical problem that mathematical psychologists had worked out by the early 1960s (Luce, Bush, & Eugene 1963). Learning occurs by remembering experienced examples, which entails that the cognitive representation of the distribution for a category is sparsely populated early on, but becomes fleshed out as experiences

accumulate, including the rare realizations of the category that define the tails of the distribution. Production is achieved by taking a random sample from the distribution (Pierrehumbert 2001). This approach can capture phenomena that are not amenable to analysis using VARBRUL. For example, Todd, Pierrehumbert, & Hay (2019) shows how an exemplar model using continuous phonetic representations can capture the very small but significant subphonemic effects relating to word frequency documented in Hay, Pierrehumbert, Walker, & LaShell (2015). However, with VARBRUL having been used longer and on more varied empirical data, exemplar theory is not yet associated with as great a range of sociophonetic observations.

The relationship between the phonological level and the morphophonological level presents the opposite set of challenges: the morphophonological level is assumed to be discrete, but it is not directly observable. If the underlying forms of the morphemes are provided (by expert linguists), then it is not too difficult to generate outcomes that include variable outcomes using variable rules. Research in Optimality Theory also showed how variable outcomes can be generated using variably ranked constraints, where an individual outcome is generated by random sampling from the ranking distributions (Anttila & Cho 1998; Boersma & Hayes 2001; see further van Oostendorp, this volume). Inferring the underlying representations if they are not known is a difficult optimization problem in languages with rich morphology. The explosive combinatorics involved in comparing each word to many other words, and the fact that computer algorithms have much worse access to semantic similarity than

people do, already mean that finding the best way to cut up complex words in a concatenative morphological system is far from trivial.

A turning point in the field was based on the observation, going back to Zipf (1936, 1949) that frequent concepts tend to be expressed using shorter forms than rare concepts. This observation suggests that human languages have a tendency towards using optimal coding, that is, towards expressing meanings in the shortest possible manner, consistent with getting messages across accurately. If we view the lexicon as codebook, we can now ask how to find the most succinct codebook that still covers the whole language. Should we include entries for *meadowlark*, *skylark*, and *woodlark*, spelled out as letters or phonemes? We might obtain shorter representations by just pointing to the forms *meadow, sky, wood, lark*, if these are needed in any case. However, *merganser* clearly needs to be spelled out, because the word does not contain any subparts that are productively used in other combinations. More generally, a succinct codebook for the lexicon can be found by identifying substrings (approximately corresponding to morphemes) that reoccur in more different words than would be predicted from their phonological form alone. This assumption, in combination with the basic grammatical assumption that affixes depend on stems, has led to surprisingly successful algorithms for segmentation of concatenative morphology (Goldsmith 2001; Creutz & Lagus 2002).

Additional challenges arise when attempting to infer morphological relationships when phonological alternations are involved. This problem can be conceptualized as the problem of optimally estimating the probabilities

of morphophonological rules (which convert one form to another) or of morphophonological relationships (in which morphologically related word forms generally match but have some points of mismatch). This inference problem is quite challenging, because not just the probabilities of the rules, but also the rules themselves, need to be inferred. Languages differ greatly in their rule inventories and many rules are not phonetically natural (Anderson 1981). A pioneering effort is Skousen (1989), which induces general rules from specific rules by iteratively collapsing rules as long as the statistical reliability is not compromised. Subsequent related efforts include Mikheev (1997); Albright & Hayes (2003); Pierrehumbert (2006). The problem continues to be actively pursued not only in linguistics, but also in statistical natural language processing in the context of efforts to engineer language processing systems for heavily inflected languages (Narasimhan, Barzilay, & Jaakkola 2015; Cotterell *et al.* 2016).

30.5 Correlations of probabilities with linguistic behaviour

A large number of experimental studies in psycholinguistics and laboratory phonology in the 1990s and 2000s found correlations between phonological probabilities as estimated from corpora, and various kinds of linguistic behaviours. Some of these studies had the goal of overturning theories that lacked probabilities, such as Chomsky & Halle (1968) and Prince & Smolensky (2008), while others had the goal of modelling language acquisition, speech production, or speech perception. Here, I review some of

the highlights of this literature.

A straightforward consequence of including probabilistic information in the phonological grammar is that combinations of rare elements or sequences are predicted to be extremely rare. In a Markov model, this prediction follows because each transition is statistically independent from the one before, and the joint probability of two independent events is the product of their individual probabilities. Any sequence that is generated by a Markov process can be assigned a probabilistic score simply by multiplying the probabilities of the transitions. (The score is normally calculated in the log domain by summing the log probabilities.) Because probabilities are by definition less than or equal to 1.0, and probabilities of 1.0 occur only in very exceptional cases, the product is almost always a smaller number than each individual probability. Models based on context-free grammars make basically the same prediction because the rules for expanding non-terminal nodes in the tree are taken to be statistically independent. Indeed, the predicted probability of a combination can be so low that it is expected to occur less than once in a corpus of realistic size. In Pierrehumbert (1994), a study of triconsonantal medial clusters in monomorphemic words of English (as in words like *palfrey* or *velcro*), this simple observation effectively explains the vast majority of missing triconsonantal clusters in a large on-line dictionary. For example, the sequence /ð.bw/ (a rare syllable coda plus a rare syllable onset) is predicted to be so rare that it would show up less than once in a large lexicon of English, and indeed monomorphemic words like /paðbwi/ do not occur. As already noted in Miller (1957), the same reasoning means that

the probabilistic score of a pseudoword gets lower and lower as we add more material. And indeed (setting aside propensities to avoid extremely short words), the likelihood that a wordform is a real word does decay rapidly with word length in every language that has been analyzed.

Well-formedness judgements of pseudowords have been shown in many experimental studies to reflect statistical scores of the forms, as determined from an analysis of the lexicon together with assumptions about the phonological grammar. Both the units that accrue probabilities, and the assumptions about how the scores for the parts contribute to the whole, differ in different studies. The simplest phonotactic score that is widely used assumes the words are generated by a bigram model defined on phonemes (instead of on letters), with scores calculated as described in the last paragraph. Scores calculated in the same way assuming a trigram model have also been found to be relevant (Bailey & Hahn 2001; Needle, Pierrehumbert, & Hay in press). As predicted by Miller (1957), assuming an n-gram model means that long words comprised of more probable parts will receive similar scores to short words with less probable parts. This prediction about the scores is reflected in human ratings. A study of word length by Frisch, Large, & Pisoni (2000) obtained ratings for pseudowords of length 2 to 4 syllables, comprised either of frequent, or rare, CV syllables. Bisyllabic words made of rare syllables were rated about the same as quadrisyllabic words made of frequent syllables. The observation is replicated in Needle, Pierrehumbert, & Hay (in press), for pseudowords ranging from 4 to 7 phonemes in length. A further important issue is how the cognitive system treats unseen sequences. Does the presence of any

unseen sequence mean that any word containing it is completely impossible? Or does the cognitive system treat unseen sequences as the limiting case of rare sequences, in effect keeping a corner of the mind open to novelty? Coleman & Pierrehumbert (1997) finds that frequent sequences can redeem completely unattested clusters. For example the pseudoword *mrupation*, combining the unattested word beginning *mr* with the common sequence *pation* is judged fairly favourably. Thus the second alternative is the correct one. In mathematical language models, smoothing methods provide a way to assign nonzero probabilities to previously unseen sequences (Jurafsky & Martin 2008). Edwards, Beckman, & Munson (2004) puts forward a different line of evidence for the relevance of empirical probabilities to the cognitive system. In a production task, they find that children make more errors in producing less common (though still legal) phonotactic sequences than in producing common ones.

While these results showed that probabilities are cognitively relevant, and that probabilistic information combines in a cumulative manner, they already contain the seed of mismatches between the cognitive system and probabilities in the strict sense. First, judgements generally correlate with log probabilities rather than probabilities per se, a point to which we return in section 30.6. Second, if we allow outputs of a Markov model to have arbitrary length, the probabilistic scores are not guaranteed to sum up to 1.0. Since probabilities do, by definition, sum up to 1.0, the probabilistic scores are not strictly speaking probabilities.

In sociophonetics, variable rules have probabilities that are associated with aspects of the context. Some of these, such as the gender or social

class of a speaker, are relatively stable over time. Others, such as formality or addressee, vary on short time scales, and individual speakers use variants with different probabilities depending on them. While such patterns of variation in production clearly demonstrate the need for a probabilistic model, they can in some cases leave considerable ambiguity about what individual speakers have learned. For example, while it is possible that people have learned to control the probabilities of hyperarticulated and reduced allophones based on the communicative situation, it is also possible that they have simply learned to speak with more effort in formal contexts and contexts that present communicative difficulties, with hyperarticulated variants more used as a consequence. This ambiguity in the interpretation lends importance to other types of evidence about the cognitive status of probabilistic information in sociophonetics. Recently, studies of sociophonetic perception have produced illuminating results. Experiments show that listeners adjust their encoding of speech depending on social information about the speaker (Johnson 2006; Sumner & Samuel 2009) and direct or indirect information about the dialectal context (Hay & Drager 2010; Sanchez, Hay, & Nilson 2015). People who are familiar with multiple dialects are also quite successful in identifying speaker dialects (Clopper & Pisoni 2004), indicating implicit knowledge of the distributions of these variants in relation to groups of speakers. These behaviours all indicate that the probabilities studied in sociophonetics reflect cognitive regularities.

Empirical probabilities are also found to be correlated with morphological decomposition and morphophonological alternations. In English, phonotactic constraints on words, in combination with the highly

productive use of compounding, mean that many complex words contain junctures that would be unlikely or impossible in simplex words. For example, the sequence /pd/ as in *topdog*, is far more likely across a word boundary than within a (monomorphemic) word, and indeed can provide a cue that the word is a compound (Daland & Pierrehumbert 2011). Hay, Pierrehumbert, & Beckman (2004*b*) show that this information is used implicitly in well-formedness judgements; judgements correlate with the probability of the single best parse. Ernestus & Baayen (2003), Albright & Hayes (2003), Pierrehumbert (2006), and Zuraw (2010) all find correlations between the empirical likelihood of alternations, and the rate at which they are applied when people are asked to produce morphological relatives of nonce words (so-called wug tests).

30.6 Probability matters vs probability matching

As Section 30.1 already noted, some probabilistic theories of phonology claim that people learn the probabilities of different phonological patterns, and reproduce them in their own outputs. Such theories are referred to as probability-matching theories. Others only make the much more general claim that probabilistic information is used in learning and reflected in adult grammars. These theories are also probabilistic, but the probabilities of different outputs are not necessarily predicted to be the same as the probabilities in the input.

The first point of view has a long history in sociolinguistics, where

probabilistic rules are used to capture variation in allophonic outcomes across dialects, speakers, and speech registers. The observation that some of this variation is remarkably persistent from one generation to the next has led some sociolinguists, notably Labov (1989), to argue that children learn the probabilities of the input they experience during language learning, and then reproduce these same probabilities. However, sociolinguists have never argued that children internalize all of the ambient probabilities. For language changes in progress, systematic differences in patterns of variation across generations are much studied. One of the main unsolved problems in sociolinguistics is to explain why some variation is relatively stable, while other variation is unstable and is resolved by changes towards a more regular system.

The claim that people internalize experienced probabilities has had a resurgence in more recent years with the rise of Bayesian models of learning. In Bayesian models, people bring a set of prior beliefs to any learning situation. In the case of phonology, these prior beliefs are expressed as probabilities associated with phonological descriptors. These descriptors are formal characterizations of sound patterns, using the ontology of the selected theory, which are either correct (true) or incorrect (false) for any individual specific example in a language. For example, one might hypothesize that people are born assuming that all words begin in consonants; or an adult second language learner might make this assumption, based on their implicit knowledge of their first language. This hypothesis can be expressed as $P(C|\#\_\_) = 1.0$ (read as 'the probability of a consonant given a word boundary is 1.0'). Accordingly, $P(V)|\#\_\_) = 0$.

This approach means that a 'prior' is a probability (or set of probabilities) assigned to the initial state of the model before the learner has any experience.[5] In the Bayesian approach, learners then update these probabilities depending on what examples they experience. As the amount of experience increases, their mental representations of the probabilities converge to the experienced probabilities. English does have words beginning in vowels, so in this hypothetical example, adult mental representations have $P(C|\#\_\_) < 1.0$ and $P(V|\#\_\_) > 0$.

In learning situations with high levels of exposure, the prior reveals itself the most at the early stages in the learning process. By the time the final state (the mature system of a fluent speaker) is reached, the amount of experience is so great that the initial state has little remaining effect. The output from a mature learner is thus predicted to display the same pattern of statistical variability as the input they experienced. Examples of recent Bayesian models of phonological learning include Shi, Griffiths, Feldman, & Sanborn (2010); Daland & Pierrehumbert (2011); Wilson & Davidson (2013); Moulin-Frier, Diard, Schwartz, & Bessière (2015). To the extent that probability-matching is observed, that is an argument that Bayesian models do a good job of capturing the critical features of the language acquisition process.

The reviews in Hayes & Londe (2006) and Hayes, Siptár, Zuraw, &

[5] Some formalizations use instead the related assumption that the initial state of the learner is instead a set of frequencies of phonological descriptors, which are accordingly counts of different types rather than probabilities of different types. However this does not affect the point being made here.

Londe (2009) claim that adult language learning is generally probability-matching (a claim that does not necessarily extend to language learning by children; see Kam & Newport 2009). Unfortunately, many of the earlier papers that are cited argue for the importance of probabilities, but not for probability-matching per se. I now summarize some of the discrepancies, because this sort of confusion appears to be widespread. To show that learning is probability-matching, it is necessary to show that the statistical patterns in the output of individual learners match those in the input, which are assumed to be the same as those in the ambient language.

One issue is that many studies only report data that has been pooled across participants. As pointed out in Estes (1956), pooled data can give a spurious appearance of probability-matching in cases where different participants learn categorically different systems. As an extreme example, if a variant occurs half the time, and some people always use it, while others never do, the pooled data will have the variant half the time. We may conclude that the variability affected the likelihood that a pattern would be learned, but not that individual learners acquired a value of $P = 0.5$. Secondly, a study in which the dependent variable (the output) is something other than the statistical patterns in the learner's linguistic productions does not by definition demonstrate probability-matching. For example, the output measures for Edwards, Beckman, & Munson (2004) are durations (in units of time) and error rates. Error rates are of course probabilities, but they are not probabilities of variants per se; instead they are probabilities describing the relationship between the intended variant and the actual variant. Pierrehumbert (1994) correlates the likelihoods of

clusters with the rank of the type frequency in the lexicon; a frequency rank is not the same thing as a probability. Bailey & Hahn (2001), Frisch, Large, & Pisoni (2000), and Hay, Pierrehumbert, & Beckman (2004*b*) all report judgements of word acceptability or typicality on a Likert scale (a scale with some number of steps, such as 5, 6 or, 7). These papers all conclude that phonology is probabilistic, but they should not be interpreted as demonstrating that phonological learning is probability-matching.

A further layer of complexity is added because in most experimental studies on the behavioural consequences of phonological probabilities, the independent variable for the statistical analysis is log scaled. Probabilisitic scores for n-gram models, as described above, are generally computed by summing log probabilities as a computational convenience. In the idealized case where people's behaviour is perfectly probability-matching, it does not matter whether both the input and the output are on a linear or a log probability scale. However, log scaling does matter when we probe the internal state of the system by looking at other dependent variables. The relationship between log frequency (as an input) and various indicative outputs (such as judgements on a Likert scale) is pretty well approximated by a straight line. This is why it is appropriate, for example, to report the correlation from a linear regression. However, a relationship that appears linear when log scaling of the independent variable is used would not be linear without the log scaling. We illustrate this with a simple example of English phonotactics. In a monomorphemic word list of English (Hay, Pierrehumbert, & Beckman 2004*b*), there are 16 words beginning in /ʃr/, 46 in /gl/, and 132 in /tr/. Thus, /#tr/ is about three times as frequent as

/#gl/, which is about three times as frequent as /#ʃr/. If all other phonotactic factors are held equal in an experiment on rating the well-formedness of pseudowords, we expect the difference in ratings between /#gl/ words and the /#tr/ words to be about the same as the difference between /#ʃr/ words and /#gl/ words. For example, if /#ʃr/ words have an average rating of 1.0, and /#tr/ words have an average rating of 7.0, then we expect the /#gl/ words to fall half way in between at ∼4.0. That's because addition on a log scale corresponds to multiplication on a linear scale, and the multiplier for both relations is ∼3. If the mental representations used a linear scale rather than a log scale, the predicted rating for the /#gl/ words is lower, at around 2.6. On a linear scale, 46 is much closer to 16 than to 132; $132 - 46 = 86$, whereas $46 - 16 = 30$. The appropriateness of log frequency scaling for the analysis of well-formedness ratings and many other types of behavioural data indicates that cognitive representations become saturated with high levels of exposure. The impact of 100 new instances of a pattern, for example, differs greatly depending on whether the learner has previously seen just one instance, or a thousand instances, of the same pattern. The same quantitative argument also becomes important for evaluating the extent of probability-matching in real-world data, where there is always some noise in the data and hence some variability around the hypothesized pattern. The relative weight of different observations in any model fitting procedure depends a lot on whether linear or log scaling of the probabilities is used, and in general log scaling appears to be more appropriate.

Log scaling is not the only way to capture nonlinearity in the

relationship between the input and the mental representations. In particular, connectionist models of language (which also appear in the literature as neural network models or Parallel Distributed Processing models) also capture such non-linearities using a variety of different equations (Feldman & Ballard 1982; Smolensky 1990; Plaut, McClelland, Seidenberg, & Patterson 1996; Goldberg 2016). However, a review of such models exceeds the scope of this chapter.

In this context, it is interesting to inquire whether people's mental representations or productions systematically also deviate from probability-matching in other ways. This possibility is already foreshadowed in Shannon's work on n-gram models. As discussed above, the model can only learn local regularities within a fixed-size window. Regularities that critically involve a larger window cannot be learned by an n-gram model using a smaller window, and will not be reproduced in the outputs. In fact, any probabilistic grammar incorporates constraints on what can be learned, just as any other type of formal grammar does. Moreton (2008) extends this observation with his discussion of cases in which cognitive biases modulate, rather than strictly determine, what can be learned. That is, controlling for the amount of statistical evidence, some kinds of generalizations are more easily learned than others, though the others might be partially learned or learned with more experience. Becker, Ketrez, & Nevins (2011) and Dawdy-Hesterberg (2014) document cases in Turkish and Arabic, respectively, in which strong statistical patterns involving vowels do not appear to be productively applied, in contrast to prior results for similar patterns involving consonants. These suggest that

the phonological encoding of consonants and vowels may differ, possibly because consonants are perceived more reliably in terms of discrete categories than vowels are (Pisoni 1975). Computational models of language variation and change point to the existence of cognitive biases towards regularity and structure, as without such biases, language systems are predicted to degenerate structurally over time. These entail that speakers do not reproduce all of the variability that they encounter, but instead develop phonological units that are distinct and discriminable. Bybee (2001), Pierrehumbert (2001), Wedel (2012), Kapatsinski (2018), and Todd, Pierrehumbert, & Hay (2019) all make proposals about how these biases figure in the perception and production systems. Kapatsinski (2018) reviews further cognitive factors that cause deviations from probability-matching behaviour, including task demands and patterns of attention. Finally, social factors shape how people attend to, encode, and remember spoken words; depending on these factors, infrequent variants may be either highlighted, or ignored, in forming long-term mental representations (Sumner, Kim, King, & McGowan 2014; Clopper, Tamati, & Pierrehumbert 2016; Todd, Pierrehumbert, & Hay 2019).

30.7 Conclusion

Probabilistic phonology has antecedents in the work of Pāṇini, and the classic distinction between accidental and systematic gaps in the lexicon is implicitly probabilistic. Mathematically explicit theories were launched with Shannon's work on information theory and Markov processes in the

1940s. Other important foundations include mathematical models of classification (Luce, Bush, & Eugene 1963) and the advances in acoustic and articulatory phonetics that made it possible to relate abstract phonological categories to observable physical parameters (Chiba & Kajiyama 1942; Fant 1970; Van den Berg 1958).

The availability of abundant recordings of conversational speech beginning in the 1970s led to two major lines of research. In sociophonetics, variable rules were developed to model effects of dialect, gender, class, and other social variables on allophonic and morphophonological variation (Sankoff & Labov 1979). Assuming an architecture in which detailed traces of linguistic experience are retained in memory, Liljencrants & Lindblom (1972) proposed a self-organizing model for the typology of vowel inventories, and Bybee and colleagues documented the importance of word frequencies in allophonic variation and diachronic phonology (Hooper 1976; Bybee 2001).

During the next decades, probabilistic phonology took on some of the major claims of mainstream generative phonology. While Chomsky & Halle (1968) and Prince & Smolensky (2008) both claimed that phonology lacks gradient and cumulative effects, systematic experiments on well-formedness judgements and other linguistic behaviours showed that such effects exist. They also showed that these effects are highly correlated with the empirical frequencies of phonological outcomes, and with the likelihoods of complex outcomes as predicted from their parts. These studies led to the conclusion that probabilities are reflected in some manner in the mental representation of phonology. The development of autosegmental-metrical phonology

provided the technical foundation for associating probabilities not merely with phonemes or allophones, but also with larger phonological units.

This era also saw significant progress on modelling the relationships amongst levels of representation in the sound structure of language. Goldsmith (2001) and Creutz & Lagus (2002) developed methods for inducing morphological decompositions from a lexicon. Skousen (1989), Mikheev (1997), and Albright & Hayes (2003) developed ways to automatically learn alternations. Models of the relationship between phonology and phonetics moved from discrete models (using fine phonetic transcriptions) to models that generate probability distributions over continuous phonetic parameters (Pierrehumbert 2001; Johnson 2005; Wedel 2012; Moulin-Frier, Diard, Schwartz, & Bessière 2015). Bayesian models of phonology provide a powerful and general account of how language learners might acquire patterns of variation from linguistic experience. However, they also predict that with high levels of experience, the learner's output will match the statistics of their input. This prediction has been evaluated in recent experimental studies. It turns out that the cognitive system distorts empirical patterns in some ways, due to cognitive biases, propensities for regularity and structure, and social influences on attention and memory. Nonetheless, it remains clear that probabilistic information is important in phonology.

References

Albright, Adam and Bruce Hayes. 2003. Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition* 90(2): 119–161.

Anderson, Stephen R. 1981. Why phonology isn't "natural". *Linguistic Inquiry* 12(4): 493–539.

Anttila, Arto and Young-mee Yu Cho. 1998. Variation and change in Optimality Theory. *Lingua* 104(1-2): 31–56.

Bailey, Todd M. and Ulrike Hahn. 2001. Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *Journal of Memory and Language* 44(4): 568–591.

Becker, Michael, Nihan Ketrez, and Andrew Nevins. 2011. The surfeit of the stimulus: Analytic biases filter lexical statistics in Turkish laryngeal alternations. *Language* 87: 84–125.

Beckman, Mary E. and Janet B. Pierrehumbert. 1986. Intonational structure in Japanese and English. *Phonology* 3: 255–309.

Van den Berg, Janwillem. 1958. Myoelastic-aerodynamic theory of voice production. *Journal of Speech, Language, and Hearing Research* 1(3): 227–244.

Boersma, Paul and Bruce Hayes. 2001. Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry* 32(1): 45–86.

Bybee, Joan. 2001. *Phonology and language use*. Cambridge University Press.

Cardona, George. 1965. On translating and formalizing Paninian rules. *Journal of the Oriental Institute (Baroda)* 14: 306–314.

Chappelier, Jean-Cédric, Martin Rajman *et al.* 1998. A generalized CYK algorithm for parsing stochastic CFG. *TAPD* 98(133-137): 5.

Charniak, Eugene. 1997. Statistical parsing with a context-free grammar and word statistics. *AAAI/IAAI* 2005(598-603): 18.

Chiba, T. and M. Kajiyama. 1942. *The vowel: Its nature and structure*. Tokyo-Kaiseikan Pub. Co., Ltd.

Chomsky, Noam. 1956. Three models for the description of language. *IRE Transactions on Information Theory* 2(3): 113–124.

Chomsky, Noam. 1957. *Syntactic structures*. MIT Press.

Chomsky, Noam and Morris Halle. 1968. *The sound pattern of English.* Harper & Row.

Clopper, Cynthia G. and David B. Pisoni. 2004. Some acoustic cues for the perceptual categorization of American English regional dialects. *Journal of Phonetics* 32(1): 111–140.

Clopper, Cynthia G., Terrin N. Tamati, and Janet B. Pierrehumbert. 2016. Variation in the strength of lexical encoding across dialects. *Journal of Phonetics* 58: 87–103.

Coleman, John and Janet B. Pierrehumbert. 1997. Stochastic phonological grammars and acceptability. In *ACL special interest group in computational phonology: Proceedings of the workshop*, 49–56. Association for Computational Linguistics.

Cotterell, Ryan, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner, and Mans Hulden. 2016. The SIGMORPHON 2016 shared task?morphological reinflection. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, 10–22.

Creutz, Mathias and Krista Lagus. 2002. Unsupervised discovery of morphemes. In *Proceedings of the ACL-02 workshop on morphological and phonological learning-volume 6*, 21–30. Association for Computational Linguistics.

Daland, R. and J. B. Pierrehumbert. 2011. Learning diphone-based segmentation. *Cognitive Science* 1: 119 – 155.

Dawdy-Hesterberg, Lisa. 2014. *The structural and statistical basis of morphological generalization in Arabic*. Northwestern University Ph.D dissertation.

Edwards, Jan, Mary E. Beckman, and Benjamin Munson. 2004. The interaction between vocabulary size and phonotactic probability effects on children's production accuracy and fluency in nonword repetition. *Journal of speech, language, and hearing research* 47(2): 421–436.

Ernestus, Mirjam Theresia Constantia and R. Harald Baayen. 2003. Predicting the unpredictable: Interpreting neutralized segments in dutch. *Language* 79(1): 5–38.

Estes, William K. 1956. The problem of inference from curves based on group data. *Psychological bulletin* 53(2): 134.

Fant, Gunnar. 1970. *Acoustic theory of speech production: With calculations based on X-ray studies of Russian articulations*. Mouton de Gruyter.

Feldman, Jerome A. and Dana H. Ballard. 1982. Connectionist models and their properties. *Cognitive Science* 6(3): 205–254.

Frisch, Stefan A., Nathan R. Large, and David B. Pisoni. 2000. Perception of wordlikeness: Effects of segment probability and length on the processing of nonwords. *Journal of Memory and Language* 42(4): 481–496.

Frisch, Stefan A., Janet B. Pierrehumbert, and Michael B. Broe. 2004. Similarity avoidance and the OCP. *Natural Language & Linguistic Theory* 22(1): 179–228.

Goldberg, Yoav. 2016. A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research* 57: 345–420.

Goldinger, Stephen D. 1998. Echoes of echoes? an episodic theory of lexical access. *Psychological Review* 105(2): 251.

Goldsmith, J. 2001. Unsupervised learning of the morphology of a natural language. *Computational Linguistics* 27(2): 154–198.

Goldsmith, John. 1990. *Autosegmental and metrical phonology*. Blackwell Publishing, Oxford, UK.

Guy, Gregory R. 1991. Explanation in variable phonology: An exponential model of morphological constraints. *Language Variation and Change* 3(1): 1–22.

Hay, J. B., J. B. Pierrehumbert, A. J. Walker, and P. J. LaShell. 2015. Tracking word frequency effects through 130 years of sound change. *Cognition* 139: 83–91.

Hay, Jennifer, Janet B. Pierrehumbert, and Mary E. Beckman. 2004*a*. Speech perception, well-formedness, and the statistics of the lexicon. *Papers in laboratory phonology VI* 58–74.

Hay, Jennifer B. and Katie Drager. 2010. Stuffed toys and speech perception. *Linguistics* 48(4): 865–892.

Hay, Jennifer B., Janet B. Pierrehumbert, and Mary E. Beckman. 2004*b*. Speech perception, well-formedness, and the statistics of the lexicon. In *Papers in Laboratory Phonology VI*, 58–74. Cambridge Univ. Press.

Hayes, Bruce and Zsuzsa Cziráky Londe. 2006. Stochastic phonological knowledge: The case of Hungarian vowel harmony. *Phonology* 23(1): 59–104.

Hayes, Bruce, Péter Siptár, Kie Zuraw, and Zsuzsa Londe. 2009. Natural and unnatural constraints in Hungarian vowel harmony. *Language* 85(4): 822–863.

Hockett, Charles F. 1958. *A course in modern linguistics*. MacMillan, New York.

Hockett, Charles F. 1960. The origin of speech. *Scientific American* 203: 88 – 111.

Hooper, Joan B. 1976. Word frequency in lexical diffusion and the source of morphophonological change. In W. C. Christie (ed.), *Current progress in historical linguistics*, 96–105. North Holland, Amsterdam.

Ito, Junko. 1988. *Syllable theory in prosodic phonology*. Garland Publishing Inc.

Johnson, Keith. 2005. Speaker normalization in speech perception. In David B. Pisoni and Robert E. Remez (eds.), *The handbook of speech perception*, 363–389. Blackwell Publishing Ltd, Malden MA.

Johnson, Keith. 2006. Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. *Journal of Phonetics* 34(4): 485–499.

Jurafsky, Daniel and James H. Martin. 2008. *Speech and language processing: An introduction to speech recognition, computational linguistics and natural language processing*. Prentice Hall.

Kam, Carla L. Hudson and Elissa L. Newport. 2009. Getting it right by getting it wrong: When learners change languages. *Cognitive Psychology* 59(1): 30–66.

Kapatsinski, Vsevolod. 2018. *Changing minds changing tools: From learning theory to language acquisition to language change*. MIT Press.

Koerner, E. F. K. 1972. Jan Baudouin de Courtenay: His place in the history of linguistic science. *Canadian Slavonic Papers* 14(4): 663–683.

Labov, William. 1989. The child as linguistic historian. *Language Variation and Change* 1(1): 85–97.

Lakoff, G. 1993. Cognitive phonology. In J. A. Goldsmith (ed.), *The last phonological rule*, 117–145. University of Chicago Press.

Liljencrants, Johan and Björn Lindblom. 1972. Numerical simulation of vowel quality systems: The role of perceptual contrast. *Language* 839–862.

Luce, Robert Duncan, Robert R. Bush, and Galanter Ed Eugene. 1963. *Handbook of mathematical psychology.* Wiley.

McCarthy, John J. 1988. Feature geometry and dependency: A review. *Phonetica* 43: 84–108.

McCarthy, John J. 1994. The phonetics and phonology of semitic pharyngeals. In Patricia Keating (ed.), *Papers in Laboratory Phonology III*, 191–283. Cambridge University Press.

Mikheev, Andrei. 1997. Automatic rule induction for unknown-word guessing. *Computational Linguistics* 23(3): 405–423.

Miller, George A. 1957. Some effects of intermittent silence. *The American Journal of Psychology* 70(2): 311–314.

Moreton, Elliott. 2008. Analytic bias and phonological typology. *Phonology* 25: 83 – 127.

Moulin-Frier, Clément, Julien Diard, Jean-Luc Schwartz, and Pierre Bessière. 2015. Cosmo (communicating about objects using sensory–motor operations): A Bayesian modeling framework for studying speech communication and the emergence of phonological systems. *Journal of Phonetics* 53: 5–41.

Narasimhan, Karthik, Regina Barzilay, and Tommi Jaakkola. 2015. An unsupervised method for uncovering morphological chains. *arXiv preprint arXiv:1503.02335* .

Needle, Jeremy, Janet B. Pierrehumbert, and Jennifer B. Hay. in press. Phonological and morphological effects in the acceptability of pseudowords. In Andrea Sims and Adam Ussishkin (eds.), *Morphological typology and linguistic cognition*. Cambridge University Press.

Nespor, Marina and Irene Vogel. 1986. *Prosodic phonology (Studies in generative grammar 28)*. Dordrecht: Foris.

Olson, Elizabeth S., Hendrikus Duifhuis, and Charles R Steele. 2012. Von Békésy and cochlear mechanics. *Hearing research* 293(1-2): 31–43.

Ostendorf, Mari and K. Ross. 1997. A multi-level model for recognition of intonation labels. In *Computing prosody*, 291–308. Springer.

Pierrehumbert, Janet B. 1994. Syllable structure and word structure: a study of triconsonantal clusters in English. In *Papers in Laboratory Phonology III: Phonological structure and phonetic form*, 168–188. Cambridge University Press.

Pierrehumbert, Janet B. 2001. Exemplar dynamics: Word frequency, lenition and contrast. In Joan L. Bybee and Paul J. Hopper (eds.), *Frequency and the emergence of linguistic structure*, 137–157. John Benjamins Publishing.

Pierrehumbert, Janet B. 2006. The statistical basis of an unnatural alternation. In *Laboratory Phonology VIII: Varieties of phonological competence*, 81–107. Mouton de Gruyter.

Pierrehumbert, Janet B. and Mary E. Beckman. 1988. *Japanese tone structure*. Cambridge, Mass.: MIT Press.

Pisoni, David B. 1975. Auditory short-term memory and vowel perception. *Memory & Cognition* 3(1): 7–18.

Plaut, David C., James L. McClelland, Mark S. Seidenberg, and Karalyn Patterson. 1996. Understanding normal and impaired word reading: computational principles in quasi-regular domains. *Psychological Review* 103(1): 56.

Potter, Ralph K., George A. Kopp, and Harriet C. Green. 1947. *Visible speech*. Van Nostrand.

Prince, Alan and Paul Smolensky. 2008. *Optimality Theory: Constraint interaction in generative grammar*. John Wiley & Sons.

Sanchez, K., Jennifer B. Hay, and E. Nilson. 2015. Contextual activation of Australia can affect New Zealanders' vowel productions. *Journal of Phonetics* 48: 76 – 95.

Sankoff, David and William Labov. 1979. On the uses of variable rules. *Language in Society* 8(2-3): 189–222.

Shannon, Claude E. 1948. A mathematical theory of communication. *Bell System Technical Journal* 27: 379–423, 623–656.

Shannon, Claude E. and Warren Weaver. 1949. *A mathematical theory of communication*. University of Illinois Press.

Shi, Lei, Thomas L. Griffiths, Naomi H. Feldman, and Adam N. Sanborn. 2010. Exemplar models as a mechanism for performing Bayesian inference. *Psychonomic Bulletin & Review* 17(4): 443–464.

Skousen, Royal. 1989. *Analogical modeling of language*. Springer Science & Business Media.

Smolensky, Paul. 1990. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence* 46(1-2): 159–216.

Sumner, M. and A. G. Samuel. 2009. The effect of experience on the perception and representation of dialect variants. *Journal of Memory and Language* 60: 487–501.

Sumner, Meghan, Seung Kyung Kim, Ed King, and Kevin B. McGowan. 2014. The socially weighted encoding of spoken words: A dual-route approach to speech perception. *Frontiers in psychology* 4: 1015.

Todd, Simon, Janet B. Pierrehumbert, and Jennifer Hay. 2019. Word frequency effects in sound change as a consequence of perceptual asymmetries: An exemplar-based model. *Cognition* 185: 1–20.

Wedel, Andrew B. 2012. Lexical contrast maintenance and the organization of sublexical systems. *Language and Cognition* 4(4): 319 – 355.

Werker, Janet F. and Richard C. Tees. 1984. Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant behavior and development* 7(1): 49–63.

Wilson, Colin and Lisa Davidson. 2013. Bayesian analysis of non-native cluster production. In *Proceedings of NELS*, vol. 40.

Zipf, G. 1936. *The psychobiology of language*. Routledge.

Zipf, G. 1949. *Human behaviour and the principle of least effort*. Addison-Wesley.

Zuraw, Kie. 2010. A model of lexical variation and the grammar with application to tagalog nasal substitution. *Natural Language & Linguistic Theory* 28(2): 417–472.